Name: _____**ANSWER KEY**_____

SID: _____

Topics in Computational Biology and Genomics
Plant & Microbial Biology / Molecular & Cell Biology / Bioengineering c146/c246

Spring 2003

MIDTERM EXAM

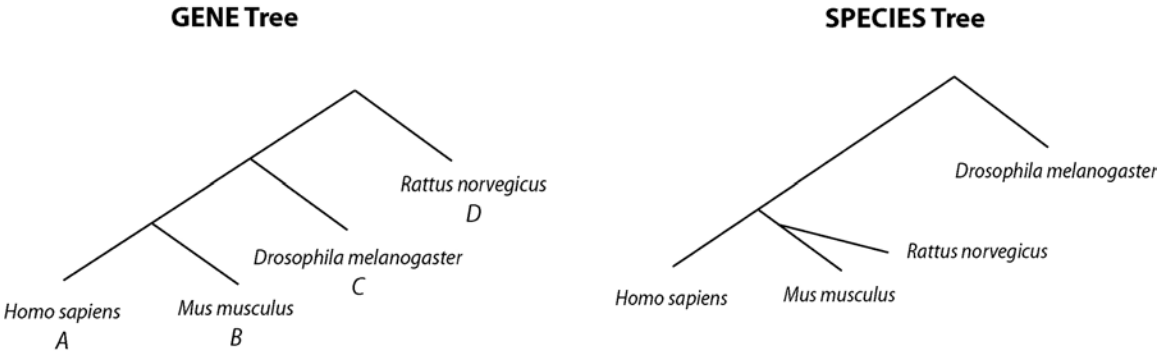
- Concise answers and clear phrases acceptable
- Complete sentences not essential
- Answers can be in functional form (*e.g.*, $\frac{1}{\sqrt{2}}$ instead of 0.707)

1	_____ /	5
2	_____ /	5
3	_____ /	10
4	_____ /	10
5	_____ /	10
6	_____ /	10
7	_____ /	10
8	_____ /	10
9	_____ /	10
10	_____ /	5
11	_____ /	15
Extra credit	_____ /	5
TOTAL	_____ /	100

NAME: _____**ANSWER KEY**_____

1. 5 points

The tree on the left depicts the evolutionary history of a family of 4 proteins found in 4 species; the tree on the right depicts the evolutionary history of the species.



(A) What is the most likely evolutionary relationship between proteins A and B? Why?

ORTHOLOGS. A and B are likely to be related by speciation, as they have very similar sequences (they are the closest together on the gene tree).

(B) What is the most likely evolutionary relationship between proteins B and D? Why?

PARALOGS. Since *Drosophila* protein C is more similar to A and B than the rat protein D, a duplication has probably occurred at the root of the gene tree. One of the paralogs has been lost in rat; the other paralog has not been found in humans, mice, or fruitfly.

2. 5 points

What is the theoretical basis that allows us to infer that proteins are orthologous from their sequences?

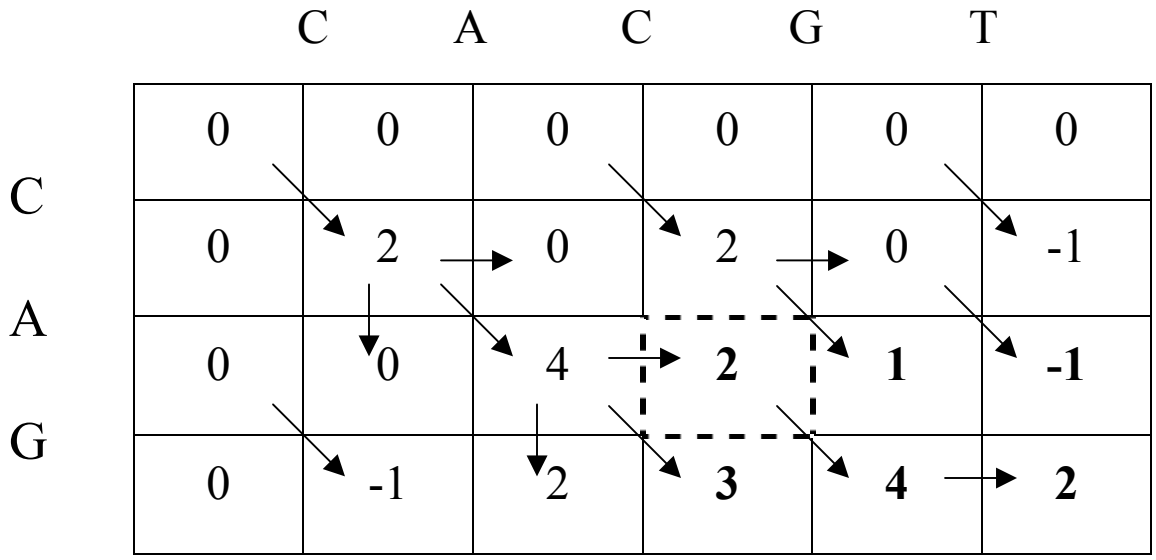
Fitch's method provides a statistical basis for inferring common ancestry (*i.e.* divergent evolution). After building a parsimony tree, sequences are inferred to be orthologous if there have been fewer inferred changes than expected by random alignments.

NAME: _____**ANSWER KEY**_____

3. 10 points

(A) (6 points) Provide the dynamic programming matrix entries and traceback vectors for the semi-global alignment (no end gap penalties for either sequence) of the sequences below. Use the following scores:

Match	+2
Mismatch	–1
Gap (fixed)	–2



(B) (2 points) What is the resulting alignment and score?

For no end gap penalties, pick the alignment with the best score anywhere on the last row or column (Score = 4).

```
CACGT
CA-G-
```

(C) (2 points) What alignment does the value in the cell with dotted borders represent?

```
CAC
CA-
```

4. 10 points

In the procedure to construct PAM matrices, how does matrix multiplication incorporate a Markov model of evolution? (*Recall*: M = A × B, where $m_{ij} = \Sigma_k \, a_{ik} \times b_{kj}$)

Each residue can mutate to every other intermediate residue after one time unit of evolution (*i.e.*, one round of matrix multiplication). In the above matrix multiplication formula, the *k* terms represent every possible intermediate residue in the transition from residue *i* to residue *j*.

5. 10 points

(A) (5 points) Write the recursion relations required for dynamic programming alignment with *generalized* gap penalties.

$A_{0,0} = 0$
$A_{0,j} = A_{i,0} = -\infty$
$A_{i,j} = \max \{ A_{i-1,j-1}\ B_{i-1,j-1}\ C_{i-1,j-1} \} + s_{i,j}$

$B_{0,0} = B_{i,0} = -\infty$
$B_{0,j} = g(j)$
$B_{i,j} = \max \{ A_{i,j-k} + g(k) \text{ for } k = 1..j$
$\quad C_{i,j-k} + g(k) \text{ for } k = 1..j$
$\quad \}$

$C_{0,0} = C_{0,j} = -\infty$
$C_{i,0} = g(i)$
$C_{i,j} = \max \{ A_{i-k,j} + g(k) \text{ for } k = 1..i$
$\quad B_{i-k,j} + g(k) \text{ for } k = 1..i$
$\quad \}$

$S_{all} = \max \{ A_{n,m}$
$\quad B_{n,m}$
$\quad C_{n,m}$
$\quad \}$

(B) (5 points) What is the time complexity (in big-oh notation) for sequence alignment with *generalized* gap penalties? for sequence alignment with affine gap penalties?

Generalized gap penalties: $O(mn^2 + nm^2)$

Affine gap penalties: $O(mn)$

6. 10 points

A BLAST database search is performed with a query sequence of length 16 (2^4) and an effective database size of 65,536 (2^{16}). The raw score of an alignment is 25, using a matrix with K = 2 and λ = 0.693 (ln 2). What is the BLAST E-value (you may leave your answer in base 2)?

$$S^* = \frac{\lambda S - \ln K}{\ln 2}$$

$$S^* = \frac{25 \ln 2 - \ln 2}{\ln 2} = 24$$

$$E = mn2^{-S^*}$$

$$= (2^{16})(2^4)(2^{-24})$$

$$= 2^{-4}$$

7. 10 points

Suppose you want to perform a database search for extremely similar hits to your query sequence. Name 3 BLAST parameters that could be changed to make the search run faster. Explain your reasons for each change in a few words.

-f (Increase)	**Score threshold**	**Increase stringency of initial word hits**
-W (Increase)	**Word size**	**Increase stringency of initial word hits**
-S (Increase)	**Cutoff score**	**Fewer hits to extend to HSP's**
-A (Decrease)	**Mutliple hits window**	**Fewer hits to extend to HSP's**

NAME: _____**ANSWER KEY**_____

8. 10 points

(A) (5 points) What are the key reasons that rigorous traditional sequence alignment by dynamic programming becomes intractable with multiple sequences?

Runs out of memory (space complexity)
Takes too much time (time complexity $O(N^L 2^L)$)

(B) (5 points) What are advantages of full dynamic programming over other multiple alignment approaches?

Guarantees optimal alignment

9. 10 points

(A) (5 points) How do CLUSTALW and CLUSTALV differ? What effects do these differences have on speed and alignment quality?

CLUSTALW weights different sequences and allows profile alignments.
Speed – No change Alignment quality – Better

(B) (5 points) How do CLUSTALW and T-COFFEE differ? What effects do these differences have on speed and alignment quality?

T-COFFEE uses local alignment information. T-COFFEE performs master-slave alignments with every other sequence to buffer against introducing incorrect gaps.
Speed – Slower Alignment quality – Better

10. 5 points

PAM matrices are constructed using the formula given to the right. What is the scaling factor, λ, associated with a PAM matrix?

$$S_{ij} = 3\log_{10}\frac{q_{ij}}{p_i p_j}$$

Note: $\log_{10} e \approx 0.434$ $\log_{10} 3 \approx 0.477$ $\log_{10} \pi \approx 0.497$

$$\log_b a = \frac{\log_{10} b}{\log_{10} a}$$

$$3\log_{10}\left(\frac{q_{ij}}{p_i p_j}\right) = \frac{\log_e \frac{q_{ij}}{p_i p_j}}{\lambda} \quad \Rightarrow \quad \lambda = \frac{\log_{10}\left(\frac{q_{ij}}{p_i p_j}\right)}{\log_{10} e} \times \frac{1}{3\log_{10}\left(\frac{q_{ij}}{p_i p_j}\right)}$$

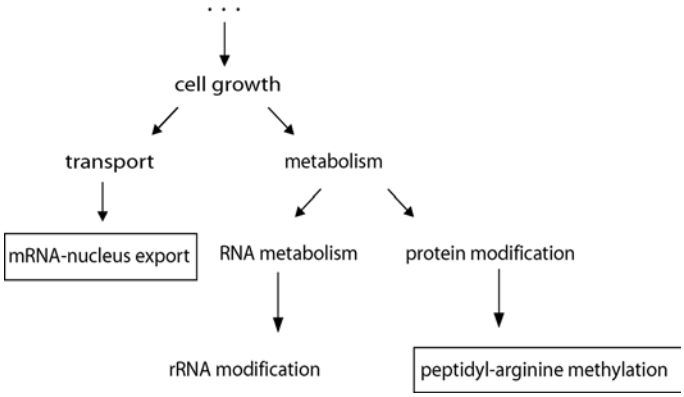
$$\lambda = \frac{1}{3\times\log_{10} e}$$

NAME: _____**ANSWER KEY**_____

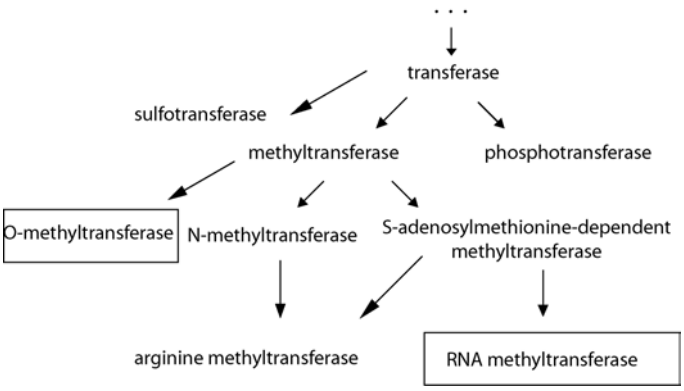
11. 15 points

A BLAST search of an unannotated query protein yielded two weak but significant hits. These hits have biochemically-verified annotations for the boxed Gene Ontology terms:

Biological Process



Molecular Function



(A) (5 points) What can you deduce about the *molecular function* of the query protein?

Methyltransferase (deepest node common to both hits)

(B) (5 points) What can you deduce about the *biological process* of the query protein?

Nothing. Biological processes may change rapidly over time, and the annotations for the hits are not very similar.

(C) (5 points) Describe a method that could help annotate the biological process of the query protein.

Phylogenetic profiles may provide a clue as to other proteins that may interact with the query protein. The query may have a similar biological role with proteins in other species that show the same patterns of conservation in different organisms.

NAME: _____**ANSWER KEY**_____

Extra credit: (5 points – all or none. Use the reverse side of the page.)
MAFFT is claimed to be a quick and accurate multiple alignment algorithm.

(A) What features of MAFFT make it quick (just saying FFT is not enough)?
(B) What features of MAFFT make it accurate?
(C) What is a fundamental limitation in its optimization function that prevents MAFFT from making a biologically optimal alignment? Why?

(A) MAFFT uses Fourier transforms to determine an offset for a well-scoring diagonal in the dynamic programming matrix

(B) MAFFT searches for amino acid residues using volume and polarity scores, which are indicators of hydrophobicity and protein structure. MAFFT also an iterative refinement method.

(C) MAFFT uses a sum of pairs scoring scheme, which fails to take into account the evolutionary history of the sequences.